Beyond bag-of-words: N-gram topic models

Presented by Marina Danilevsky
Motivation: topical phrases

• Topical phrases are more informative
  o “time” vs “time series”, “frequent” vs “frequent pattern mining”

• Phrase meaning may not always be derivable from its constituent unigrams
  o “compact disc” is a music medium, but neither unigram (“compact” or “disc”) can successfully convey this concept – need the entire bigram

• A phrase may be meaningful in one context and not another:
  o “white house” is a meaningful phrase in a political topic; but
  o “white house” is not a meaningful phrase in a real estate topic

• A phrase may have different meanings in multiple contexts:
  o “compact disc” is a music medium in a music topic; and
  o “compact disc” is a small region bounded by a circle in a math topic
Talk Outline

• Bigram Topic Model (BTM)
• LDA Colocation Model (LDACOL)
• Topical N-gram Model (TNG)

• Brief intro to Hierarchical Pitman-Yor Processes

• Phrase-Discovering LDA (PDLDA)
  o Experiment & Results: LDA vs TNG vs PDLDA
<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>number of topics</td>
</tr>
<tr>
<td>$D$</td>
<td>number of documents</td>
</tr>
<tr>
<td>$W$</td>
<td>number of unique words</td>
</tr>
<tr>
<td>$N_d$</td>
<td>number of word tokens in document $d$</td>
</tr>
<tr>
<td>$z_{i}^{(d)}$</td>
<td>the topic associated with the $i^{th}$ token in the document $d$</td>
</tr>
<tr>
<td>$x_{i}^{(d)}$</td>
<td>the bigram status between the $(i-1)^{th}$ token and $i^{th}$ token in the document $d$</td>
</tr>
<tr>
<td>$w_{i}^{(d)}$</td>
<td>the $i^{th}$ token in document $d$</td>
</tr>
<tr>
<td>$\theta^{(d)}$</td>
<td>the multinomial (Discrete) distribution of topics w.r.t. the document $d$</td>
</tr>
<tr>
<td>$\phi_{z}$</td>
<td>the multinomial (Discrete) unigram distribution of words w.r.t. topic $z$</td>
</tr>
<tr>
<td>$\psi_{v}$</td>
<td>in Figure 1(b), the binomial (Bernoulli) distribution of status variables w.r.t. previous word $v$</td>
</tr>
<tr>
<td>$\psi_{zv}$</td>
<td>in Figure 1(c), the binomial (Bernoulli) distribution of status variables w.r.t. previous topic $z$/word $v$</td>
</tr>
<tr>
<td>$\sigma_{zv}$</td>
<td>in Figure 1(a) and (c), the multinomial (Discrete) bigram distribution of words w.r.t. topic $z$/word $v$</td>
</tr>
<tr>
<td>$\sigma_{v}$</td>
<td>in Figure 1(b), the multinomial (Discrete) bigram distribution of words w.r.t. previous word $v$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Dirichlet prior of $\theta$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Dirichlet prior of $\phi$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Dirichlet prior of $\psi$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Dirichlet prior of $\sigma$</td>
</tr>
</tbody>
</table>
1. draw Discrete distributions $\sigma_{zw}$ from a Dirichlet prior $\delta$ for each topic $z$ and each word $w$;

2. for each document $d$, draw a Discrete distribution $\theta^{(d)}$ from a Dirichlet prior $\alpha$; then for each word $w^{(d)}_i$ in document $d$:

   (a) draw $z^{(d)}_i$ from Discrete $\theta^{(d)}$; and

   (b) draw $w^{(d)}_i$ from Discrete $\sigma_{z^{(d)}_i w^{(d)}_{i-1}}$.
Can decide whether to generate a unigram or a bigram. But, a discovered bigram is always a bigram, regardless of context. So “white house” would remain a phrase in a document about real estate.

Topical N-gram Model

Topic assignments for the two terms in a bigram are not required to be identical. Post-hoc, assign topic of phrase to be the topic of the last unigram (or the most popular).

Setting every $x_i=1$ gives BTM. Making $\sigma$ conditioned on previous word only gives LDACOL.

Brief Intro to Hierarchical Pitman-Yor Processes

N-gram Language Models and Smoothing

- (N-1)th order Markov model:

\[ P(\text{sentence}) = \prod_{i} P(\text{word}_i|\text{word}_{i-N+1} \ldots \text{word}_{i-1}) \]

- Large vocabulary size means naively estimating parameters of this model from data counts is problematic for N>2.

\[ P_{\text{ML}}(\text{word}_i|\text{word}_{i-N+1} \ldots \text{word}_{i-1}) = \frac{C(\text{word}_{i-N+1} \ldots \text{word}_i)}{C(\text{word}_{i-N+1} \ldots \text{word}_{i-1})} \]

- Naïve priors/regularization fail as well: most parameters have no associated data.
  - Smoothing.

http://www.gatsby.ucl.ac.uk/~ywteh/research/compling/bayesnll.pdf
Smoothing on Context Tree

- *Context* of conditional probabilities naturally organized using a tree.
  \[ P^{\text{smooth}}(\text{road}|\text{south parks}) = \lambda(3)Q_3(\text{road}|\text{south parks}) + \lambda(2)Q_2(\text{road}|\text{parks}) + \lambda(1)Q_1(\text{road}|\emptyset) \]

- Smoothing makes conditional probabilities of neighbouring contexts more similar.

- Later words in context more important in predicting next word.

http://www.gatsby.ucl.ac.uk/~ywteh/research/compling/bayesnll.pdf
Hierarchical Bayesian Models on Context Tree

- Parametrize the conditional probabilities of Markov model:
  \[ P(\text{word}_i = w | \text{word}_{i-N+1}^i = u) = G_u(w) \]
  \[ G_u = [G_u(w)]_{w \in \text{vocabulary}} \]
- \( G_u \) is a probability vector associated with context \( u \).

What is \( P(G_u | G_{pa(u)}) \)?

http://www.gatsby.ucl.ac.uk/~ywteh/research/compling/bayesnll.pdf
Pitman-Yor processes produce distributions over words given by a power law distribution with index 1+d.

- Small number of common word types
- Large number of rare word types.

http://www.gatsby.ucl.ac.uk/~ywteh/research/compling/bayesnll.pdf
Chinese Restaurant Process: Pitman-Yor

- Generative Process:

\[
p(\text{sit at table } k) = \frac{c_k - d}{\theta + \sum_{j=1}^{K} c_j}
\]

\[
p(\text{sit at new table}) = \frac{\theta + dK}{\theta + \sum_{j=1}^{K} c_j}
\]

- Defines an exchangeable stochastic process over sequences \(x_1, x_2, \ldots\)

- The de Finetti measure is the Pitman-Yor process,

\[
G \sim PY(\theta, d, H)
\]

\[
x_i \sim G \quad i = 1, 2, \ldots
\]

http://www.gatsby.ucl.ac.uk/~ywteh/research/compling/bayesnll.pdf
When $d=0$, Pitman-Yor reduces to Dirichlet

- **Generative Process:**

  $p(\text{sit at table } k) = \frac{c_k - d}{\theta + \sum_{j=1}^{K} c_j}$

  $p(\text{sit at new table}) = \frac{\theta + dK}{\theta + \sum_{j=1}^{K} c_j}$

- Defines an exchangeable stochastic process over sequences $x_1, x_2, ...$

- The de Finetti measure is the **Dirichlet process** $G \sim \text{DP}(\theta, H)$

  $x_i \sim G \quad i = 1, 2, ...$

*Pitman-Yor Processes are also known as Two-parameter Poisson-Dirichlet Processes*

http://www.gatsby.ucl.ac.uk/~ywteh/research/compling/bayesnll.pdf
Chinese Restaurant Process: Pitman-Yor

- customers = word tokens.
- H = dictionary.
- tables = dictionary lookup.

Dictionary look-up sequence:

```
cat, dog, cat, mouse
```

Word token sequence:

```
cat, dog, dog, dog, cat, dog, cat, mouse, mouse
```
Hierarchical Pitman-Yor Language Models

- Parametrize the conditional probabilities of Markov model:

$$P(\text{word}_i = w | \text{word}_{i-N+1}^i = u) = G_u(w)$$

$$G_u = [G_u(w)]_{w \in \text{vocabulary}}$$

- $G_u$ is a probability vector associated with context $u$.

- Place Pitman-Yor process prior on each $G_u$.  

Diagram:

- $G_{\emptyset}$
- $G_{\text{parks}}$
- $G_{\text{south parks}}$
- $G_{\text{along south parks}}$
- $G_{\text{at south parks}}$
- $G_{\text{to parks}}$
- $G_{\text{university parks}}$
Phrase-Discovering LDA Model (PDLDA) using HPYP

Topic assignments for the two terms in a bigram are not required to be identical. Post-hoc, assign topic of phrase to be the topic of the last unigram (or the most popular).

Each phrase is drawn from one topic
• \(u\) is a context vector consisting of the phrase topic and the past \(m\) words: \(u = <z_i, w_{i-1}, \ldots, w_{i-m}>\)
• \(\pi(u)\) denotes the prefix of \(u\), the vector with the rightmost element of \(u\) removed
• \(|u|\) denotes the length of \(u\)
• \(\theta\) represents an empty context.
**PDLDA vs TNG**

**TNG**
- Topic assignments for terms in a bigram aren’t required to be identical
- Post-hoc, assign topic of phrase to be the topic of the last unigram
- Observing a bigram under one topic does not make it more likely under another topic or make its constituent unigrams more probable.

**PDLDA**
- Each phrase is drawn from one topic. Each token is drawn from a distribution conditioned on its context $u$
  - $u$ is a context vector consisting of the phrase topic and the past $m$ words: $u = \langle z_i, w_{i-1}, \ldots, w_{i-m} \rangle$
  - When $m = 1$, this conditioning is analogous to TNG’s word distribution
  - But, PDLDA’s word distributions used are Pitman-Yor processes (PYPs) linked together into a tree structure
  - Hierarchical construction creates smoothing among different contexts
    - Phrases can share probability mass between contexts, don’t need to independently infer the probability of every bigram under every topic
    - Advantageous with a smaller corpora, or with many topics
Hierarchical Pitman-Yor Toy Example

- Illustration of HPYP for a toy two-word vocabulary $V = \{\text{honda}, \text{civic}\}$ and two-topic ($T = 2$) model with $m = 1$
- Each node $G$ in the tree is a PYP whose base distribution is its parent node
- $H$ is chosen to be a uniform distribution over $V$: $H(w) = 1/|V|$
- E.g.: when the context is $u = z_1 : \text{honda}$, the bold path is followed and the probability of the next word is calculated from the shaded node by combining predictions from all the nodes along the bold path.
• The HPYP gives the following probability for a word following the context $u$ being $w$:

$$P_u(w \mid \tau, a, b) = \frac{c_{uw} - a_{u|t_{uw}}}{b_{|u|} + c_{u.}} + \frac{b_{|u|} + a_{u|t_{uw}}}{b_{|u|} + c_{u.}} P_{\pi(u)}(w \mid \tau, a, b)$$

• $P_{\pi(\theta)}(w \mid \tau, a, b) = G_{\theta}(w)$  
  ($\theta$ represents an empty context)

• $c_{uw.}$ is the number of customers eating dish $w$ in restaurant $u$

• $t_{uw}$ is the number of tables serving $w$ in restaurant $u$
  
  • $t_{uw}$ represents the current seating arrangement

• Dots indicate marginal counts: e.g., $c_{uw.} = \sum_k c_{uwk}$ where $c_{uwk}$ is the number of customers eating $w$ in restaurant $u$ at table $k$. 
Phrase Intrusion Detection Experiment

Dataset: subset of the TREC AP corpus (2,246 documents)
• One of the one hundred topics found by PDLDA on the Touchstone Applied Science Associates (TASA) corpus (Landauer and Dumais, 1997).
• Each column within a box shows the top fifteen phrases for a topic and is restricted to phrases of a minimum length of one, two, or three words, respectively.
• The rows are ordered by likelihood.
Comments

• This type of output can itself be mined
  o We could transform each topic into a single list of mixed-length phrases by applying topical ranking function
• All these methods assume a phrase is a sequence of consecutive unigrams
• Can not model phrases as concepts:
  o Cannot identify “mining frequent pattern” and “frequent pattern mining” as the same concept
  o Cannot identify that “mining frequent pattern” concept is present in “mining top-k frequent closed pattern”
Thank you!
Questions?